Principal Coordinate Maps of Molecular Potential Energy Surfaces

OREN M. BECKER

Department of Chemical Physics, School of Chemistry, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel

Received 12 January 1998; accepted 11 March 1998

ABSTRACT: Obtaining useful representations of molecular conformation spaces and visualizing the associated potential energy surfaces is a complex task, mainly due to the high dimensionality of these spaces. Principal component analysis (PCA), which projects multidimensional data on low-dimensional subspaces, is thus becoming a common technique for studying such spaces. Three issues, relating to the use of principal component techniques for mapping molecular potential energy surfaces, are discussed in this study: the effectiveness of the projection; its accuracy; and the mapping procedure. The effectiveness of PCA is demonstrated through detailed analyses of principal component projections of several peptides. In these cases PCA projected conformation space into a subspace smaller even than that defined by the peptide's backbone dihedral angles. The average accuracy as well as the distribution of errors in the projection (i.e., the errors in reproducing individual distances) are studied as a function of the dimensionality of the projection. The wide variation in accuracy between different systems suggests that it is imperative to indicate the accuracy of the projection whenever PCA projections are used. Furthermore, when projecting potential energy surfaces on the principal two-dimensional (2D) plane, the projection errors result in artificial roughening of the surface. A new mapping procedure, the ''minimal energy envelope'' procedure, is introduced to overcome this problem. This procedure yields relatively smooth ''energy landscapes,'' which highlight the basin structure of the real multidimensional energy surface. It is demonstrated that the projected potential energy maps can be used for charting conformational transitions or dynamic trajectories in the system. J Comput Chem 19: 1255–1267, 1998

Keywords: conformation space; potential energy surface; energy landscape; principal component analysis; principal coordinate analysis; peptide

Correspondence to: O. M. Becker; becker@sapphire.tau.ac.il; http://www.tau.ac.il/~becker

Contract/grant sponsor: Tel Aviv University

CCC 0192-8651 / 98 / 111255-13

Introduction

Mesoscopic systems with many degrees of freedom include liquids, glasses, and macromolecules. These systems are under intensive study in physics, chemistry, and biology. In all cases the thermodynamic and dynamic properties of the system are determined by the nature of the potential energy surfaces (also known as ''energy landscapes'').[1] It is the potential energy surface that controls processes such as protein folding[2,3] or glass transitions.[4,5] Analyses of molecular conformation spaces, over which the potential energy surfaces are defined, are used for locating stable structures of potential drug molecules,[6,7] for analyzing molecular flexibility[8] and in the context of ligand docking.[9,10] These applications are typically based on detailed analyses of large samples of molecular conformations, generated with methods such as quenching from high-temperature molecular dynamics trajectories[8,11,12] or Monte Carlo sampling followed by minimization.[8,13,14]

Potential energy surfaces (PES) can be characterized by their minima, which correspond to locally stable configurations, and by transition regions connecting the minima. In small systems, which have only few minima, it is possible to use a direct approach and describe the entire potential energy surface. For systems with many degrees of freedom and a very large number of minima, a direct approach to the PES becomes very difficult. In recent years several new methods were introduced for accurate in-depth analyses of complex potential energy surfaces. For example, low resolution representations of the PES of two small proteins was obtained by Brooks and collaborators.[15,16] Becker et al.[17,18] used ''topological analysis'' to obtain the overall topographies of the PES of several peptides, whereas Berry et al.[19–21] analyzed specific connectivity pathways to characterize the basin structure on the PES of several clusters. Another method developed recently for analyzing potential energy surfaces of clusters is the Lid method of Schön, Sibani, and others.[22,23]

A related, although separate, question is that of visualizing the multidimensional potential energy surfaces. In general, creating useful representations of molecular conformation spaces (and potential energy surfaces) is complicated by the high dimensionality of these spaces. A molecule of N atom has $3N$ degrees of freedom, and its corresponding conformation space is $3N - 6$ dimensional. As a result, even relatively small molecules have very large conformation spaces. Thus, methods such as principal component analysis, which project multidimensional data on low-dimensional subspaces, are very suitable for representing and visualizing conformation spaces and molecular dynamics trajectories that traverse these spaces.[24–26] Principal component analysis (PCA) was introduced to protein simulations under the name ''quasiharmonic analysis'' by Karplus et al. and was found useful for comparing normal modes to molecular dynamics results.[30,31] Recently, Berendsen et al.[32–35] have used PCA to describe protein ''essential dynamics,'' by focusing on the subspace in which most of the atomic motion occurs during dynamics. The question regarding the robustness of using the ''essential'' subspace as a basis for efficient dynamic simulations (as opposed to an analysis tool) is not fully resolved.[34,36]

In recent years, it has become increasingly more common to use PCA projection techniques for visualizing molecular dynamics trajectories.[13,26,37–39] In this application, the trajectory is projected on principal directions (or principal planes), thus making dynamic transition between basins visible. PCA is also used for visualizing broad conformation samples in the context of conformational analysis.[33,39–41] In a few cases, the potential energy surfaces themselves were projected and visualized in the reduced subspaces.[39,41]

In general, two related ''principal component analysis'' techniques are used: PCA and principal coordinate analysis (PCoorA). Both methods project the $n \times m$ data matrix, $\mathbf{M}$, a distribution of n points in an m variable space, on a transformed axes set in which a low-dimensional subspace, containing most of the relational information about the original distribution, can be identified. In the context of conformational analysis the starting point is a set of n conformations described by n points $P_i(q_{i1}, \ldots, q_{im})$ in an m-dimensional conformation space. Each matrix element, M_{ij}, is equal to q_{ij}, the jth coordinate of the ith conformation. From this starting point, PCA and PCoorA follow different routes.

Principal component analysis (PCA) takes the m-coordinate vectors $\mathbf{q}$ associated with the n conformation sample and calculates the square $m \times m$ $\mathbf{M}^{\mathbf{T}}\mathbf{M}$ matrix, reflecting the relationships between the *coordinates*. This matrix, also known as the

covariance matrix, **C**, is defined as:

$$\mathbf{C} = \left\langle (\mathbf{q} - \langle \mathbf{q} \rangle)(\mathbf{q} - \langle \mathbf{q} \rangle)^T \right\rangle \quad (1)$$

where the averaging is over the conformation sample (in Cartesian space $m = 3N$ for an N atom molecule). The covariance matrix, **C**, is diagonalized to obtain the eigenvectors that capture most of the variation in atom positional fluctuations.

PCoorA,[42] on the other hand, operates on the square $n \times n$ $\mathbf{MM^T}$ matrix, reflecting the relationships between the *conformations*. The elements of this matrix, also known as the distance matrix, Δ, are the distances between two conformations P_i and P_j:

$$d_{ij}^2 = \sum_{k=1}^{m} (M_{ik} - M_{jk})^2 \quad (2)$$

where it is tacitly assumed that the coordinates of the different conformations were first overlaid one on top of the other in an optimal way. Because the distance, d_{ij}, can also be obtained from the $n \times n$ matrix, **A**, of latent roots (eigenvectors), one can use this matrix for the projection, defining $A_{ij} = -1/2 d_{ij}^2$ and $A_{ii} = 0$ for $(i, j = 1, 2 \ldots n)$.[42] To guarantee that matrix **A** has a zero root (and thus guarantee that it corresponds to a real configuration) it is ''centered,'' so that the sum of every row and of every column of **A** is zero. This centering, which does not alter the distances, d_{ij}, is defined as:

$$A_{ij}^* = A_{ij} - \langle A_{ij} \rangle_i - \langle A_{ij} \rangle_j + 2\langle A_{ij} \rangle_{ij} \quad (3)$$

where $\langle \ldots \rangle_k$ is the mean over all specific indices $k = i, j, ij$. The centered matrix $\mathbf{A}^*$ is diagonalized using standard matrix algebra to obtain the latent eigenvectors and the diagonal matrix of eigenvalues. The resulting eigenvalues (normalized) give the percentage of the projection of the original distribution on the new coordinate set, and the eigenvectors (scaled by their corresponding eigenvalues) give the new coordinates of the original distribution on the new coordinate set, and the eigenvectors (scaled by their corresponding eigenvalues) give the new coordinates of the original points in the new axes frame.

In his original work, Gower[42] showed the duality between principal coordinate analysis (PCoorA, denoted by him as Q techniques) and principal component analysis (PCA, denoted by him as R techniques). Given a specific multivariate sample, *both* methods give the *same* analysis results; that is, the same eigenvectors and eigenvalues. Technically, when $n \leq m$ (i.e., when there are more coordinates than conformations), PCoorA is computationally more efficient because it involves diagonalizing a smaller matrix. When $n > m$ PCA is advantageous from the computational point of view. However, on today's computers, and for the type of molecular systems studied so far (up to several thousand conformations and/or coordinates), such considerations are largley irrelevant. Nonetheless, even when $n > m$, PCoorA may be a better choice for visualization because it leads directly to the coordinates of all the points in the newly transformed principal space (the mapped coordinates).

As discussed earlier, PCA techniques are used for visualizing molecular dynamics trajectories, molecular samples, and even molecular potential energy surfaces. Despite the increasing popularity in studying molecular systems, a detailed analysis of the quality of these projections (level of dimensionality reduction, average errors, distribution of errors) was not reported. These issues are addressed in this article through a detailed study of PCA projections of several peptide systems.

Effective Dimensionality

The main motivation for using either principal coordinate analysis (PCoorA) or principal component analysis (PCA) is to construct a low-dimensional representation of the original high-dimensional data. The notion behind this approach is that the *effective* (or *essential* as some call it[32]) dimensionality of the studied systems (in this case molecular conformation spaces) is significantly smaller than their full dimensionality ($3N - 6$ degrees of freedom for an N atom molecule). Indeed, applications of PCoorA to polypeptides have shown that the effective dimensionality of the studied conformation spaces is significantly smaller than the dimensionality of the full space. Examples of molecular systems studied in this way are the proteins crambin[39] and BPTI[38] and the peptides Met-enkephalin,[13] isobutyryl-$(ala)_3$-NH-methyl,[40,41] hexapeptide analogs of $(ala)_6$,[18] and several septa-peptide RGD analogs (this work). A similar reduction of dimensionality was obtained when analyzing molecular dynamics trajectories by PCA (e.g., refs. 32 and 37). Nonetheless, the accuracy of the projection method is not often discussed.

In principal component projections the resulting eigenvalues represent the variation of the original distribution along the principal directions. When the eigenvalues and corresponding eigenvectors are sorted in decreasing order the first eigenvector represents the axis of maximal variance, the second is the axis with the second largest variance, and so forth. Projection of the distribution onto the first two or three dimensions represents the best possible planar and 3D projections of the distribution. The accumulative sum of the first *s*-normalized eigenvalues represents the average quality of the representation when projecting the original distribution into an *s*-dimensional subspace (see next section). When the accumulative sum equals 100% it means that the d_{ij} distances are represented without distortion in the projection subspace.

Figure 1 shows the accumulative sum of the normalized eigenvalues obtained from PCoorA of four different peptides. The first is the blocked peptide isobutyryl-(ala)$_3$-NH-methyl (IAN), the second is a linear septapeptide GRGDSPC with charged N- and C-terminals (RGD), the third is the same GRGDSPC peptide with uncharged terminals (RGDu), and the fourth is a backbone cyclized analog of GRGDSPC (cycRGD). The last three peptides represent analogs of interest, including the RGD (Arg—Gly—Asp) sequence, which has an important biological activity through its involvement in cell adhesion[43] and, as such, has been a target for extensive studies.[43,44] The PCoorA analysis of IAN, based on the all-atom root mean square (rms) distance of all 139 known local minima,[17,45] has been discussed in length elsewhere.[40] The PCoorA of the three RGD peptides is based on backbone rms distances of a sample of 500 local minima generated by refining (cooling to 300 K and minimizing) transient conformations along 500 ps of molecular dynamics at 1000 K. The trajectories were calculated with the program CHARMM,[46] using a distance-dependent dielectric constant. The details of the sampling procedure and evidence that the procedure results in an extensive sampling of the peptide conformation spaces are given elsewhere.[18]

Figure 1 shows that all four peptides share similar projection profiles. The accumulative sum of eigenvalues reaches 55–85% within two dimensions, 70–90% within three dimensions, and all reach over 90% within ten dimensions. This means that, on average, the pairwise distances projected

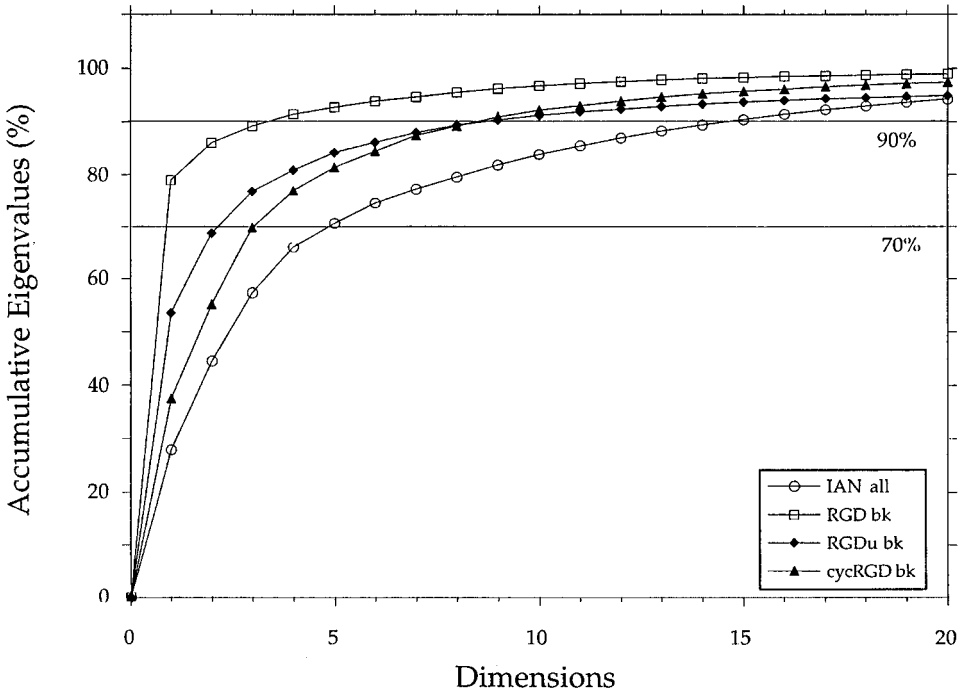

FIGURE 1. The accumulative sum of normalized eigenvalues from PCoorA of four peptides: isobutyryl-(ala)$_3$-NH-methyl (IAN); GRGDSPC with charged terminals (RGD); GRGDSPC with uncharged terminals (RGDu); and cyclic CGRGDSPC with unchanged terminals (cycRGD). The RGD analogs PCoorA are based on backbone rms distances; the IAN PCoorA is based on all-atom distances.

into the best three-dimensional (3D) subspace deviates from the actual distances by as little as 10–30%. Namely, relatively accurate 3D visual representations of the whole conformation space (to the extent that the sample is representative) can be obtained. For certain purposes, such as charting 3D energy surfaces, even the 2D projection gives a realistic (although somewhat distorted) geometrical projection.[41]

To assess the usefulness of these projections the number of *effective* PCA dimension required for an accurate representation should be compared with other dimensionality assessments. In particular, it should be compared with the full dimensionality of each molecule and to the dimensionality of its dihedral angle representation. The full dimensionality of $3N - 6$ degrees of freedom per N atom molecule, although correct, is expected to be significantly larger than the effective dimensionality because it includes many ''unimportant'' coordinates such as those associated with fluctuations of hydrogen atoms. A more realistic estimate of the expected effective dimensionality is the dimensionality of the dihedral angles subspace, and especially the subspace of backbone ϕ, ψ dihedral angles. PCA should be considered effective if the dimensionality obtained from it is smaller than the dimensionality of the dihedral angles subspace, because only then it does focus on a smaller effective subspace. This, in turn, depends on the desired level of accuracy set for the accumulative normalized eigenvalues. The dimensionality associated with a 70% accurate projection is, of course, smaller than that associated with a projection accurate to 90%.

Table I compares the calculated dimensionality of the aforementioned four peptides (full space and dihedral space) with the effective dimensionality obtained through PCA at different accuracy levels. It is seen that, at 70% accuracy, the PCA projections focus on very small effective subspaces, typically those of three-dimensions or fewer, making it a very powerful projection method at this level of accuracy. At 90% accuracy the advantages of the PCA projections lessens and becomes case-sensitive. At this level of accuracy, depending on the individual system studied, the effective subspaces obtained by PCA are smaller than or comparable to the most restricted estimate of dimensionality—that of backbone dihedral angles. PCA projections are better when based on backbone rms distances (the RGD analogs) or on distances measured in the reduced dihedral angle subspace (IAN), compared with projections based on all-atom distances. This suggests that, when less ''noise'' is introduced into the distance matrix (as with the aforementioned distance measures) it is easier for the projection algorithm to focus on an effective subspace. Similar results have been reported for a 0.9-ns trajectory of lysozyme.[32] PCA of that system's C_α covariance matrix resulted in 20 eigenvectors (out of 387), which recovered 90% of the total motion. However, the same level of motion required 35 eigenvectors (out of 3792) when using an all-atom covariance matrix.

The theory of PCoorA, just outlined, is designed to ensure that all the eigenvalues will be positive and that a real configuration is obtained in the projected space (by centering the **A** matrix). However, it turns out that in many PCoorA studies there are some eigenvectors that are negative (indicating imaginary values for the coordinates along these axes), which reflect some non-Euclidean properties of the spaces.[47] In the backbone PCoorA of the aforementioned four peptides we found that the accumulative magnitude of the negative eigen-

TABLE I.
PCoorA for Different Peptides: Expected and Observed Effective Dimensionality at Different Levels of Accuracy.[a]

	IAN (all)	IAN (dihed.)	RGD (all)	RGD (bk.)	RGDu (bk.)	cycRGD (bk.)
Full space	78	10	267	63	63	63
All dihedrals	10	10	38	21	21	21
Backbone (ϕ, ψ)	7	7	14	14	14	14
70% accuracy	5	2	3	1	2	3
90% accuracy	15	4	19	4	9	10
95% accuracy	22	6	33	7	20	15

[a]Notation: all, PCoorA based on all-atom rms distances in Cartesian coordinates; bk., PCoorA based on backbone rms distances in Cartesian coordinates; dihed, PCoorA based on rms distances in dihedral angle space.[38]

values was no more than 3% of the accumulative sum of all positive eigenvalues (in RGD and cycRGD as little as 0.4%). When calculating IAN distances in dihedral angle space, the non-Euclidean effects are naturally stronger and the percent of negative eigenvalues is 8%.[40] In other systems, we sometimes find that this percentage can reach up to 10%, in accordance with other reported results.[13]

PCA projections, in general, are sensitive to outlying points. The largest distance between sample points will have a large contribution to the orientation of the first principal direction, regardless of whether this point is of physical significance.[38] An example is the initial conformation used in the sampling procedure, which is often arbitrary (and of higher energy), and not really relevant for the studied conformation space. One easy way to prevent this ''distortion'' is to filter out high-energy conformations (or the first few conformations along the trajectory) prior to the PCoorA (unless these conformations are also of interest). Somewhat related is the spiral distortion reported by Troyer and Cohen when projecting a BPTI trajectory onto the 2D principal plane, which in their case contained about 60% of the variance.[38] This distortion is related to mapping the high-dimensional space onto a low-dimensional space and becomes less significant as more dimensions are included in the projected low-dimensional subspace.

Quality of Projection

The eigenvalues in PCoorA and PCA identify and select the ''best'' set of effective coordinates with which to represent the system. In general, this means, selecting the s-dimensional subspace in which the point-to-point distances, d_{ij}, are best reproduced. This results in the best s-dimensional projection available for the given multivariate data. In this section, we explore the quality of the PCoorA projection as a function of the subspace dimensionality s.

In PCoorA, the normalized eigenvalues, λ_i (the latent roots), are directly related to the average deviation of the projected distances, calculated in an s-dimensional subspace $d_{ij}^{(s)2}$, from the exact distances, d_{ij}^2. This is due to the fact that each eigenvector is scaled so that its sum of squares equals the corresponding latent root, $\mathbf{v}^T\mathbf{v} = \lambda$. This means that, if λ_k is small, the contribution $(Q_{ik} - Q_{jk})^2$ to the distance between points P_i and P_j will also be small, where Q_{ik} is the coordinate of the ith conformation along the kth principal direction. The calculated distance between points P_i and P_j in a s-dimensional subspace is given by:

$$d_{ij}^{(s)2} = \sum_{k=1}^{s} (Q_{ik} - Q_{jk})^2 \qquad (4)$$

As a consequence, the quality of the projection, defined as the average deviation of the distances in s dimensions, $d_{ij}^{(s)2}$, from the exact distances, d_{ij}^2, is given by the sum of the first s eigenvalues:

$$1 - \sum_{k=1}^{s} \lambda_k = \langle d_{ij}^2 - d_{ij}^{(s)2} \rangle_{ij} \qquad (5)$$

where $\langle \ldots \rangle_{ij}$ is the average over all possible ij distances in the ensemble. By definition, when $s = m$ (the full dimensionality) the average deviation equals zero. For unisotropic data, a small number of dimensions is sufficient for reproducing distances with relatively small average deviations.

Eq. (5) reflects only the average of all projected distances, but it does not answer the question of regarding the reliability of the representation of individual distances; that is, the question of the distribution of *errors* in the projection. Figure 2 shows two distributions of deviation from exact distances (i.e., errors in the representation) as a function of dimensionality. Figure 2a shows the distribution of errors for IAN when using PCoorA projections based on all-atom rms and Figure 2b shows the same distribution for RGD using PCoorA projection based on backbone rms. As expected, in both cases, the average deviation decreases as the number of dimensions in the projection increases. Figure 2 also shows that the distribution of deviations (errors) changes as a function of dimensionality. The very broad distribution at 1D (which means that the individual distances are poorly represented in one-dimension) becomes narrower as the number of dimensions increases. It is also seen that in low dimensions almost all of the distances are underestimated, whereas, with higher dimensional projections some distances are overestimated and some are underestimated, bringing the average closer to zero. In the case of the RGD peptide (Fig. 2b) the one- and two-dimensional projections show a bimodal distribution, which becomes unimodal at higher dimensions.

Figure 3 summarizes the distribution of deviations from exact distances (i.e., distribution of errors) as a function of the dimensionality of the projection. The standard deviation of the distribu-

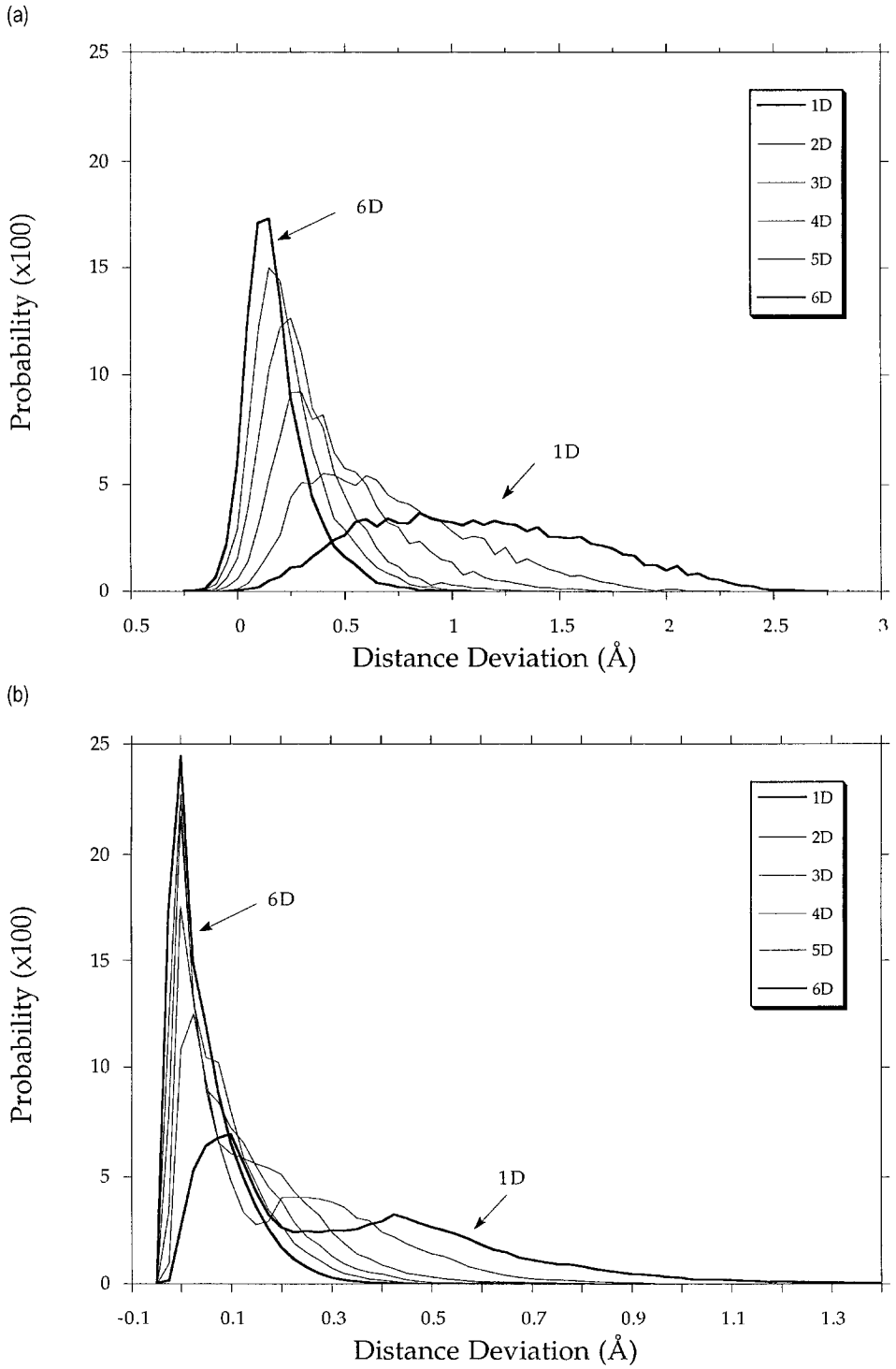

FIGURE 2. Distributions of deviations of the $d_{ij}^{(s)}$ distances calculated in an *s*-dimensional principal coordinate subspaces from the exact distances, d_{ij}, calculated in the full space (i.e., distribution of errors in the projection), for $s = 1 \ldots 6$. (a) IAN using PCoorA projections based on all-atom rms; (b) RGD using PCoorA projections based on backbone rms.

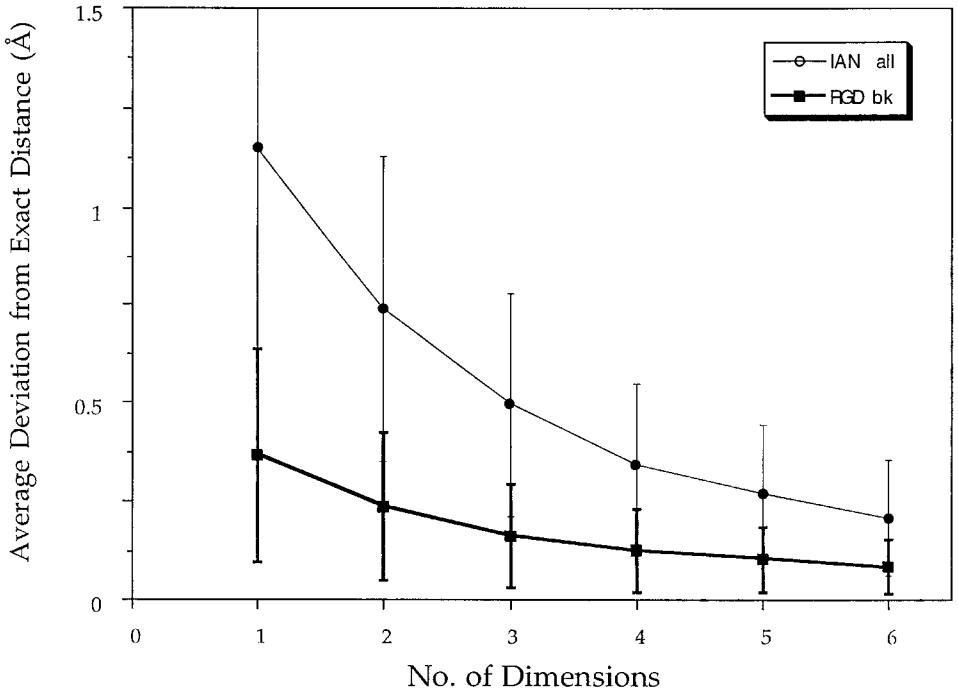

FIGURE 3. The average and standard deviation (depicted as error bars) of the distribution of the deviations of the *s*-dimensional distances $d_{ij}^{(s)}$ from the exact distances, d_{ij}, as a function of the dimensionality (s = 1 . . . 6) for the two systems shown in Figure 2.

tion is drawn as an error bar around the average deviation. For example, in IAN's PCoorA, based on all-atom rms, the 1D projection has an average deviation of 1.15 ± 0.51 Å, but the 3D projection already improves to an average deviation of 0.49 ± 0.28 Å and to 0.20 ± 0.15 Å at the 6D projection. These numbers should be compared with the average of the distances in this system, 2.14 ± 0.50 Å (maximal distance 3.37 Å).

To conclude, the wide variation in accuracy between the different systems and as a function of dimensionality indicates that, whenever PCA projections are used, it is imperative to at least indicate the average accuracy of the projection; that is, the sum of normalized eigenvalues that correspond to the principal axes participating in the projection.

"Minimum Energy Envelope" Mapping of Potential Energy Surfaces

Principal component analysis is a useful framework for visualizing the multidimensional molecular potential energy surface (PES), as was recently demonstrated by Becker,[41] who constructed the potential energy surface of the IAN tetrapeptide. However, the potential energy surface plotted in that work required substantial analysis of the topology of the surface, including detailed knowledge of the barriers on the PES, information rarely available in standard conformational analyses. A cruder, but still very informative, representation of the potential energy surface can be obtained by charting an energy contour plot, based on the potential energies of all sample points, projected on the first two or three principal coordinates. Such maps, which lack barrier information, should clearly be regarded with some suspicion. It is well understood that minima may sometimes be near in terms of the main principal coordinates yet still be separated by very high barriers. This point was clearly analyzed and demonstrated by Becker[40] in the case of IAN tetrapeptide. Thus, in the absence of barrier information (which is the common situation), it is expected that PCA representations will at most reflect the *overall* shape of the potential surface (limited by the average error of the projection), although not necessarily its details. The lack of information on the barriers is somewhat compensated by the presence of "empty spaces," which correspond to poorly sampled regions, often asso-

ciated with high energy (barrier) regions. Such a 2D energy contour map was constructed for crambin by Caves et al.[39]

The presently discussed procedures construct PES maps from energies of *all* sampled conformations. However, energy maps constructed from all of the projected points are, by definition, sensitive to projection errors. These errors can introduce significant roughness into the map, to the extent of obscuring the surface topography. Two neighboring points in the projection are indeed close in conformation space, but they are not necessarily exact neighbors. Thus, despite their nearness, their energies can be quite different causing the projected energy contour map to be rough. When the number of sampling points is small this problem can be overcome by a simple smoothing procedure, as used in mapping the energy landscape of IAN.[41] When more conformations are involved smoothing is problematic. Here we suggest an alternative solution, which reduces the roughness by using the *minimum energy envelope* procedure for visualizing complex multidimensional energy surfaces using principal coordinates.

Figure 4 shows the energies of the 500 RGD conformations, described in the previous section, projected on the first and second principal axis (calculated based on backbone rms distances). In this case, the first principal axis holds 75% of total variance and the second principal axis holds 9% of total variance. Figure 4a shows that the energy landscape of this molecule is divided along the first principal coordinate into two completely separate regions, both spatially and energetically. For clarity, the two conformation clusters are marked by different symbols. The first region, which is to the right of the origin, includes 64.6% of the conformations in the sample, and is associated with the global minimum (of the conformation sample) containing mainly local minima with energies below -150 kcal/mol. The second region, which is to the left of the origin, includes 35.4% of the conformation sample and is associated with a group of higher energy minima, all with energies above -160 kcal/mol. Figure 4b adds a second dimension to the same two conformation clusters, showing that the cluster of low energy conformation is narrow in both principal axes, whereas the cluster of higher energy conformations has a broader profile. Even though our data do not include barrier information it is reasonable to assume that the two regions are separated by relatively high barriers (indicated by the unsampled ''empty'' region between them).

Figure 4 shows that, in many instances, there is more than one conformation at a given coordinate value along the first two principal axes, resulting in a broad spread of energy. As discussed earlier, this spread is associated with projecting higher principal dimensions onto the 2D principal plane (these dimensions are less important than the first two, but still not negligible). Plotting an energy contour map based on all the conformations in the sample will thus result in a very rough surface. On the other hand the ''minimum energy envelope'' curves, schematically represented by the dashed curves in both frames of Figure 4, highlights the underlying basin structure. The justification for these curves is that they filter out projection errors introduced due to ''flattening'' the higher dimensions in principal coordinate space. Thus, a smoother potential energy surface is obtained by plotting a surface based on the lowest energy in each $(\Delta X, \Delta Y)$ interval along the axes, highlighting the underlying basin structure. Figure 5 shows the ''minimum energy envelope'' potential energy surface of the RGD peptide. The two basins on the surface are clearly seen. Of course, the resulting map will depend on the actual values chosen for the $(\Delta X, \Delta Y)$ intervals over which the ''envelope'' surface is calculated. As the grid mesh size is made finer the resulting surface becomes more detailed, but also rougher.

Figure 6 shows the shape of the RGD conformations characteristic of the two basins seen in Figure 5, proving that the two basins are indeed structurally distinct. Figure 6a depicts the lowest energy conformation that belongs to the deep basin and Figure 6b shows the conformation associated with the lowest energy in the shallower basin. The energy of the second conformation is 19.36 kcal/mol higher than the global minimum (-178.10 kcal/mol and -158.74 kcal/mol). The main difference between the two conformations (and hence between the two conformation clusters) is that the low energy conformation is folded forming a strong electrostatic interaction between the charged C- and N-terminals. This interaction is missing from the conformations in the high-energy cluster.

The potential energy surface, projected on principal coordinates using the ''minimum envelope'' procedure, in spite of the fact that they do not directly contain barrier information, can be used as frameworks for future studies of the molecular system. For example, molecular dynamic trajectories, or least energy pathways, calculated independently of the cartographic process, can be charted

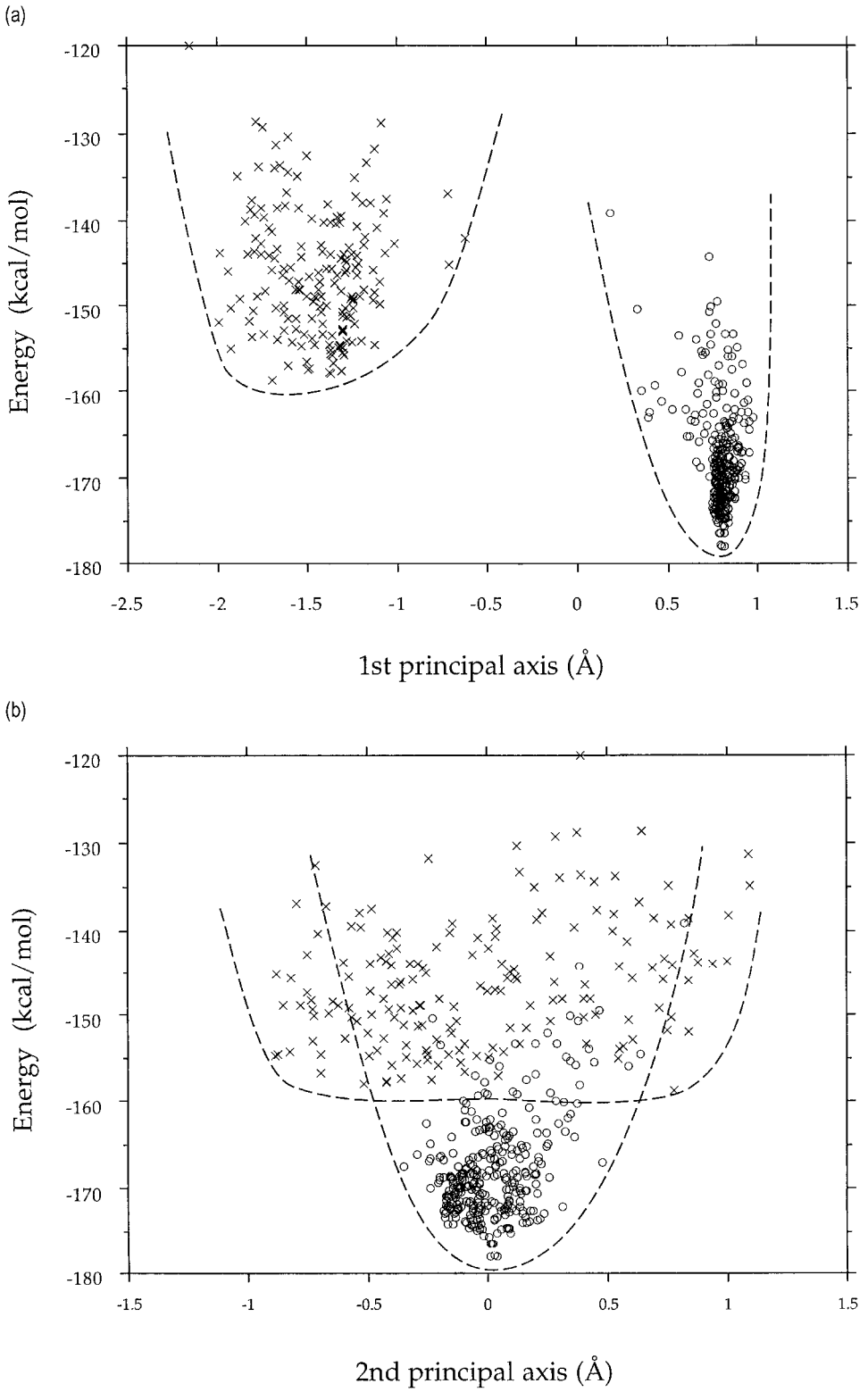

FIGURE 4. The energies of 500 RGD conformations projected on (a) the first and (b) second principal axis (based on backbone rms distances). The two conformation clusters are marked by different symbols. The dashed curves schematically represent the ''minimum energy envelope'' that underlines the two regions.

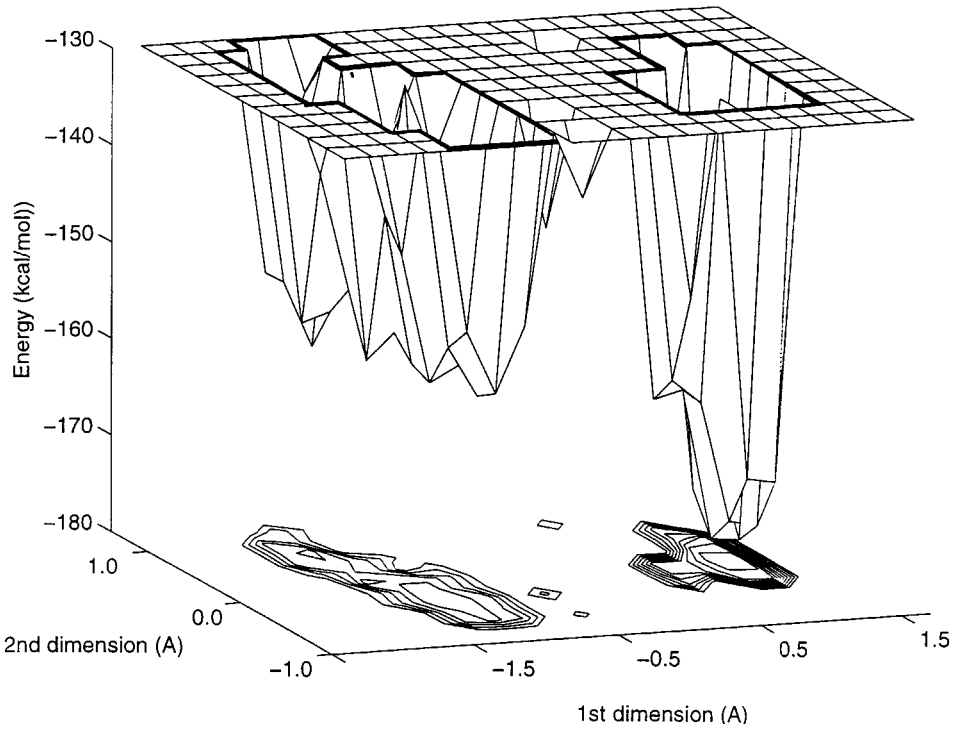

FIGURE 5. A ''minimum energy envelope'' mapping of the potential energy surface of the RGD peptide ($\Delta x = \Delta y =$ 0.2 Å). The two basins on the potential energy surface are seen.

(i.e., projected) on these maps. To map additional conformations on a precalculated plane of maximum variance we use the procedure developed by Gower.[47] The new structure, s, must be placed in the principal coordinate space in a way that preserves its distances from the M reference conformations used for generating the map. Given the set of M distances d_{si} $(i = 1 \ldots M)$ between the new point s and the reference points, the coordinate Q_{sj} of point s along the jth principal axis of

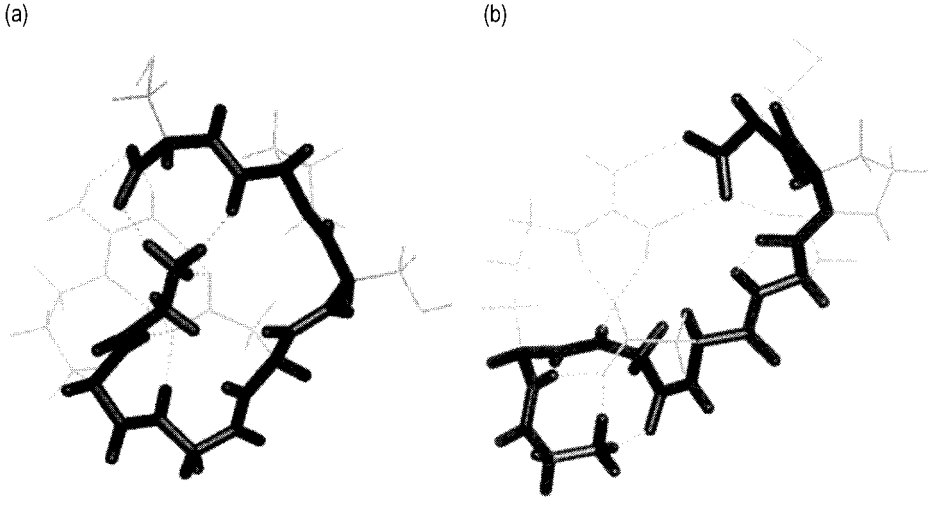

FIGURE 6. Two RGD conformations: (a) the lowest energy conformation and (b) the conformation associated of the lowest energy in the high-energy cluster (Fig. 6), which is 19.36 kcal / mol higher than the lowest energy conformation. The low-energy conformation is folded to form an electrostatic interaction between its charged C- and N-terminals.

the system is given by:

$$Q_{sj} = \frac{1}{2\lambda_j} \sum_{i=1}^{M} Q_{ij}(A_{ii}^{*} - d_{si}) \qquad (6)$$

where Q_{ij} is the coordinate of the ith reference point along the jth principal axis, A_{ii}^{*} is defined by eq. (3) and equals the square of the distance of point i from the centroid, and λ_j is the jth eigenvalue. This procedure was recently used by Abagyan and Argos[13] for projecting molecular dynamics trajectories of Met-enkephalin on a 2D principal plane constructed from eight reference conformations.

Figure 7 shows a projection of the least-energy path connecting the lowest energy conformations of each basin (Fig. 6) on the 2D contour map of the energy surface of RGD shown in Figure 5. The least-energy path was calculated using the ''conjugate peak refinement'' algorithm.[48] The maximal barrier along this path is 90.3 kcal/mol higher than the energy of the minimum of the shallow basin (109.7 kcal/mol higher than the deepest minimum) and is located closer to the shallow basin than to the deep one. The complicated structure of the projected path indicates contribution from other principal direction (mainly from the third principal axis). These can be resolved by a 3D projection or by additional 2D plots (e.g., the projection on the plane made from the first and third principal axes).

To conclude, we have demonstrated how PCA projections together with the ''minimum energy envelope'' procedure yield smooth surfaces that quantitatively represent the energy landscape of large complex molecules. Clearly, these landscapes are accurate only in a coarse way, revealing the overall structure of the surface but not necessarily its finer details. In such large systems, this coarseness may be considered an advantage. The smoothing of finer details results both from the lack of specific barrier information (only very large barrier region are represented as unsampled regions) and from the ''envelope'' smoothing procedure (which compensates for the missing dimensions in the projection). In principle, by adding barrier information to the projection maps, more accurate and better detailed landscapes can be obtained (see, e.g., the energy landscape of the IAN tetrapeptide).[41]

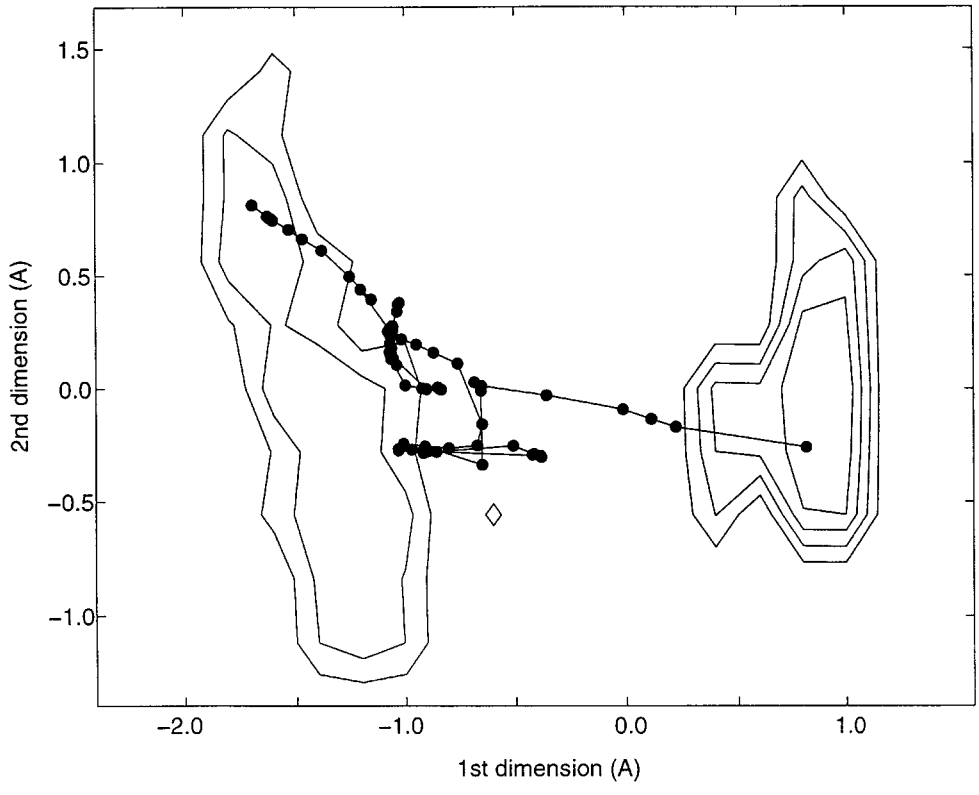

FIGURE 7. A projection of the least energy path connecting the lowest energy conformations of each basin (Fig. 6) on the 2D map of the potential energy surface of the RGD peptide shown in Figure 5.

Acknowledgments

I thank Leo S. D. Caves for many helpful discussions.

References

1. E. R. Davidson, *Chem. Rev.* **93**, 2337 (1993).
2. M. Karplus and E. Shakhnovich, In *Protein Folding*, T. E. Creighton, Eds., W. H. Freeman, New York, 1992, p. 127.
3. J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, *Proteins*, **21**, 167 (1995).
4. P. W. Anderson, B. I. Halperin, and C. M. Varma, *Philos. Mag*. **25**, 1 (1972).
5. J. L. Green, J. Fan, and C. A. Angell, *J. Phys. Chem.*, **98**, 13780 (1994).
6. A. E. Howard and P. A. Kollman, *J. Med. Chem.*, **31**, 1669 (1988).
7. A. R. Leach, In *Reviews in Computational Chemistry*, *Vol. 2*, K. B. Lipkowitz and D. B. Boyd, Eds., VCH, New York, 1991, p. 1.
8. J. C. Hempel, R. M. Fine, M. Hassan, W. Ghoul, A. Guaragna, S. C. Koerber, Z. Li, and A. T. Hagler, *Biopolymers*, **36**, 282 (1995).
9. I. D. Kuntz, *Science*, **257**, 1078 (1992).
10. I. D. Kuntz, E. C. Meng, and B. K. Shoichet, *Acc. Chem. Res.* **27**, 117 (1994).
11. F. H. Stillinger and T. A. Weber, *Phys. Rev.*, **A28**, 2408 (1983).
12. R. E. Bruccoleri and M. Karplus, *Biopolymers*, **29**, 1847 (1990).
13. R. Abagyan and P. Argos, *J. Mol. Biol.*, **225**, 519 (1992).
14. Z. Li and H. A. Scheraga, *Proc. Natl. Acad. Sci. USA*, **84**, 6611 (1987).
15. E. M. Boczko and C. L. Brooks III, *Science*, **269**, 393 (1995).
16. F. B. Sheinerman and C. L. Brooks III, *Proc. Natl. Acad. Sci. USA*, **95**, 1562 (1958).
17. O. M. Becker and M. Karplus, *J. Chem. Phys.*, **106**, 1495 (1997).
18. Y. Levy and O. M. Becker, *Phys. Rev. Lett*. (submitted).
19. R. S. Berry and R. E. Kunz, *Phys. Rev. Lett.*, **74**, 3951 (1995).
20. K. D. Ball, R. S. Berry, R. E. Kunz, F-Y. Li, A. Proykova, and D. J. Wales, *Science*, **271**, 963 (1996).
21. R. E Kunz and R. S. Berry, *J. Chem. Phys.*, **103**, 1904 (1995).
22. J. C. Schön, *Ber. Bunsenges. Phys. Chem.*, **100**, 1388 (1996).
23. P. Sibani et al., *Europhys. Lett.*, **22**, 479 (1993).
24. D. A. Case, *Curr. Opin. Struct. Biol.*, **4**, 285 (1994).
25. B. B. Brooks et al., *J. Comput. Chem.*, **16**, 1522 (1995).
26. A. N. E. García, *Phys. Rev. Lett.*, **68**, 2696 (1992).
27. M. Karplus and J. N. Jushick, *Macromolecules*, **14**, 325 (1981).
28. T. Ichiye and M. Karplus, *Proteins*, **11**, 205 (1991).
29. D. Perahia, R. M. Levy, and M. Karplus, *Biopolymers*, **29**, 645 (1990).
30. S. Hayward, A. Kitao, and N. Gô, *Prot. Sci.*, **3**, 936 (1994).
31. S. Hayward, A. Kitao, and N. Gô, *Proteins*, **23**, 177 (1995).
32. A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, *Proteins*, **17**, 412 (1993).
33. A. Amadei, A. B. M. Linssen, B. L. de Groot, D. M. F. van Alten, and H. J. C. Berendsen, *J. Biomol. Struct. Dynam.* **13**, 615 (1996).
34. B. L. de Groot, A. Amadei, D. M. F. van Alten, and H. J. C. Berendsen, *J. Biomol. Struct. Dynam.*, **13**, 741 (1996).
35. D. M. F. van Alten, A. Amadei, A. B. M. Linssen, V. G. H. Eijsink, G. Vriend, and H. J. C. Berendsen, *Proteins*, **22**, 45 (1995).
36. M. A. Balsera, W. Wriggers, Y. Oono, and K. Schulten, *J. Phys. Chem.*, **100**, 2567 (1996).
37. A. E. García, In *Nonlinear Excitations in Biomolecules*, M. Peyrard, Ed., Springer, Berlin, 1994, p. 191.
38. J. M. Troyer and F. E. Cohen, *Proteins*, **23**, 97 (1995).
39. L. S. D. Caves, J. D. Evanseck, and M. Karplus *Prot. Sci.*, **7**, 649 (1998).
40. O. M. Becker, *Proteins*, **27**, 213 (1997).
41. O. M. Becker, *J. Mol. Struct.* (*Theochem*) **398**–**399**, 507 (1997).
42. J. C. Gower, *Biometrika*, **53**, 325 (1996).
43. E. Ruoslahti and M. D. Pierschbacher, *Science*, **238**, 491 (1987).
44. M. D. Pierschbacher and E. Ruoslahti, *J. Biol. Chem.*, **262**, 17294 (1987).
45. R. Czerminski and R. Elber, *J. Chem. Phys.*, **92**, 5580 (1990).
46. B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplys, *J. Comput. Chem.* **4**, 187 (1983).
47. J. C. Gower, *Biometrika*, **55**, 582 (1968).
48. S. Fischer and M. Karplus, *Chem. Phys. Lett.*, **194**, 252 (1992).